Behavioral analysis with movement cluster model for concurrent actions

Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2011, Article ID 365307, 21 pages
doi:10.1155/2011/365307

Research Article

Behavioural Analysis with Movement Cluster Model for Concurrent Actions

Zsolt L. Husz, Andrew M. Wallace (EURASIP Member), and Patrick R. Green

Joint Research Institute of Signal and Image Processing, Heriot-Watt University, Edinburgh EH14 4AS, UK

Correspondence should be addressed to Andrew M. Wallace, a.m.wallace@hw.ac.uk

Received 1 April 2010; Revised 27 August 2010; Accepted 20 September 2010

Academic Editor: Dan Schonfeld

We present an approach to model articulated human movements and to analyse their behavioural semantics. First, we describe a novel dynamic and behavioural model that uses movements, a sequence of consecutive poses, from motion captured video data to establish priors for both tracking and behavioural analysis. Second, using that model, we show how we can both learn and subsequently recognise human activity. Activities are modelled and recognised independently to allow concurrent and complex actions. Finally, we combine activity recognition with tracking to produce an overall evaluation of the effectiveness of the approach using publicly available datasets.

1. Introduction

The analysis of human activity and behaviour from video sequences, often termed *video analytics*, has applications in surveillance, entertainment, intelligent domiciles, and medical diagnosis. Human tracking [1] is the first phase of most video analytic systems, and indeed, we have previously described a hierarchical particle filter to recover torso and limb positions and movements from video sequences [2]. However, in this paper, we focus on the behavioural understanding of human movements or transitions between poses.

Video analytics requires inferences on human behaviour, but this can vary widely in complexity from simple actions (walk, run, jump, etc.) to more complex deductions on intent or interaction (shoplifting or holding a conversation). Turaga et al. [3] provides an overview of recent algorithms to classify such a range of behaviours, some classified as simple motion patterns, the *actions*, and some as more complex, frequently multiperson *activities*, which may be composed from simple actions. For example, Ma et al. [4] extract motion trajectories from human subjects (considered as single entities) and can infer interactive behaviour (fight, chase) from the relative motion in comparison with learnt examples. Similarly, Li et al. [5] use the motion trajectories of football players to determine the "plays" of activities of the football groups from analysis of these combined tracks. The assumption in these works is that the movement of a single entity, a human blob, is sufficient to describe a singular motion, and that more complex behaviour can be inferred from interactions between trajectories. The tactical arrangement of a football team is an obvious example of this. However, more complex human activities may require more detailed analysis of the structure within the human movement, as for example using limb positions and movements in our work, and indeed in other work such as that of Deutscher and Reid [6]. However, whereas limbs are the building blocks of the human form, it is also possible to analyse behaviour on the basis of more abstract features, such as the extremities of a contour used by Yu and Aggarwal [7], or the signed distance transform used by Nater et al. [8]. These latter examples are effectively forms of 2D blob analysis for action recognition and anomaly detection, that echo the early days of binary computer vision using extracted human silhouettes.

The novelty of this paper is that we describe complex behavioural analysis that infers multiple conclusions using a *Movement Cluster Model* (MCM). This generalises both global actions (i.e., full body defined) such as walking and running, and more detailed actions (i.e., body part defined) such as forward arm movement. To achieve this, we use sequences of human pose parameters extracted from video data.

The work we describe has some parallels with manifold analysis by Lui et al. [9] and with Hierarchical Gaussian Process Latent Variable Models (HGPLVM) [10–13]. A HGPLVM is similar to an MCM in that it decomposes the pose-space into hierarchical subspaces. However, the key difference between our model and that of the previous examples is that rather than use static poses, these are substituted within an MCM with movements, that is sequences of poses.

Returning to the classification of Turaga et al. [3], the MCM belongs to the class of *parametric action detection* algorithms. This is the most complex class of action recognition, followed by activity recognition algorithms. We use movement, action and activity, defined in Section 2.2, as our three abstraction levels of perception, similar to Bobick [14]. In [14], a movement is a continuous motion characterised by the trajectory in some configuration space. An activity is a statistical temporal combination of movements, and if understood within a context, it becomes an action. İkizler and Forsyth [15] used acts and activities as basic blocks of recognition. Acts, similar to the actions above, are expressed as *Hidden Markov Models* (HMM) with a low number of states. Similar states are interconnected in a larger HMM representing activities. Similarly, Green and Guan [16] have four abstraction levels: the dyname, skill, activity and context. The first three corresponded to movement, action and activity while the context is a prior knowledge about the conditional probabilities of the skills.

The MCM is introduced in Section 2. The MCM is then trained and used for motion prediction in Section 3, while Section 4 evaluates the semantic analysis on motion captured (MOCAP) data. In Section 5 we present results of behavioural analysis on video sequences using tracked data from the HPPF [2] tracker incorporating MCM movement prediction. In Section 6 we compare and contrast our results with those of other workers on the basis of the published results. Finally we conclude our paper in Section 7.

2. Behaviour Analysis and the Movement Cluster Model

Using the *Movement Cluster Model* (MCM), we learn, simulate and subsequently analyse human movement and behaviour by extracting common poses and transitions between those poses, leading to definitions of activities and actions that represent similar movements and transitions using a clustering approach.

2.1. Human Model and Tracking. Analysis may use direct image features [17, 18] or extracted pose features [7, 13]. For MCM, the input is extracted pose sequences of the Articulated Hierarchical Human Model (AHHM) [2, 19]. A Pose Vector (PV) is the set of body configuration parameters that completely define a pose. For an articulated model, this is the torso position and the set of joint angles of the limbs. Fixed parameters, such as body part shape and size, are constant for an individual, and are therefore not included into the PV.

A Body Feature Vector (BFV) is a partition ϕ of the PV $\mathrm{p}, \mathrm{b} = \mathrm{p}^{\phi}$. It is a subset of joint angle, body position and orientation parameters. While a pose is completely and uniquely defined by a PV, several BFVs prescribed by a partition ϕ exist for a pose. This is important, since models for individual body parts (e.g., lower leg, arm, etc.) or for the whole body can be built from the appropriate partition and its BFV.

The principal source of AHHM data in this paper is the verification set of the HumanEva dataset. Later, we use AHHM data acquired by applying an articulated body tracker to video sequences, the Hierarchical Partitioned Particle Filter (HPPF) [2, 19]. The HPPF recovers the probability density of an AHHM, and each pose hypothesis, a.k.a. particle, is a PV. Figure 1 shows some poses recovered by the HPPF, as wireframe human bodies on top of the input image sequence.

However, in the rest of this paper, we consider that a tracking algorithm is already in place, and focus on the semantic analysis of the AHHM model. The behavioural analysis should be independent of the prior tracking, and therefore the HPPF may be replaced with any of the recent articulated human tracking algorithms such as those of Qu and Schonfeld [20], Bo and Sminchisescu [21], Raskin et al. [12], Rius et al. [22], and Taylor et al. [23].

2.2. The Movement Cluster Model. With no intentional content, a *movement* is a short, continuous sequence of poses, which are represented as a BFV. For example,

$$\mathrm{m} = \left[\mathrm{b}_{l_m-1}, \ldots, \mathrm{b}_1, \mathrm{b}_0\right] \tag{1}$$

is a sequence of consecutive poses, represented by a vector of the current b_0, and previous $\mathrm{b}_1, \ldots, \mathrm{b}_{l_m-1}$. l_m is the length or duration of the movement. An *action* is a short sequence of poses (e.g., *leg rising*, *arm still*). It is usually, but not exclusively, defined by one or more body parts. An action and a movement are similar, but actions have an intentional content that can be described semantically by a label λ. An *activity* is a symbolic characterization of the body over a limited time, bearing an intention.

A *Movement Cluster* (MC) is a set of similar movements. MCs are the building elements of actions, and an MC can be part of multiple actions. If a movement is part of an MC then the statistical probability of the MC being an Action A results in the probability of movement being Action A.

To make an analogy with language, activities are sentences, actions are words, movements are letters, and BFVs are the sequences of curves forming the letters. This structure is shown in Figures 2 and 3. Like previous authors [14–16, 24], we build semantic knowledge from simpler towards complex structures. The translation of the PV (e.g., the tracking data) to symbolic description is performed through MCs.

The MCM consists of an MC set and the probability of transitions between them. The MCM can be used to predict movement or PVs, and as such to improve tracking with prior knowledge. Further, if an MC has an associated probability distribution of semantic action labels, then

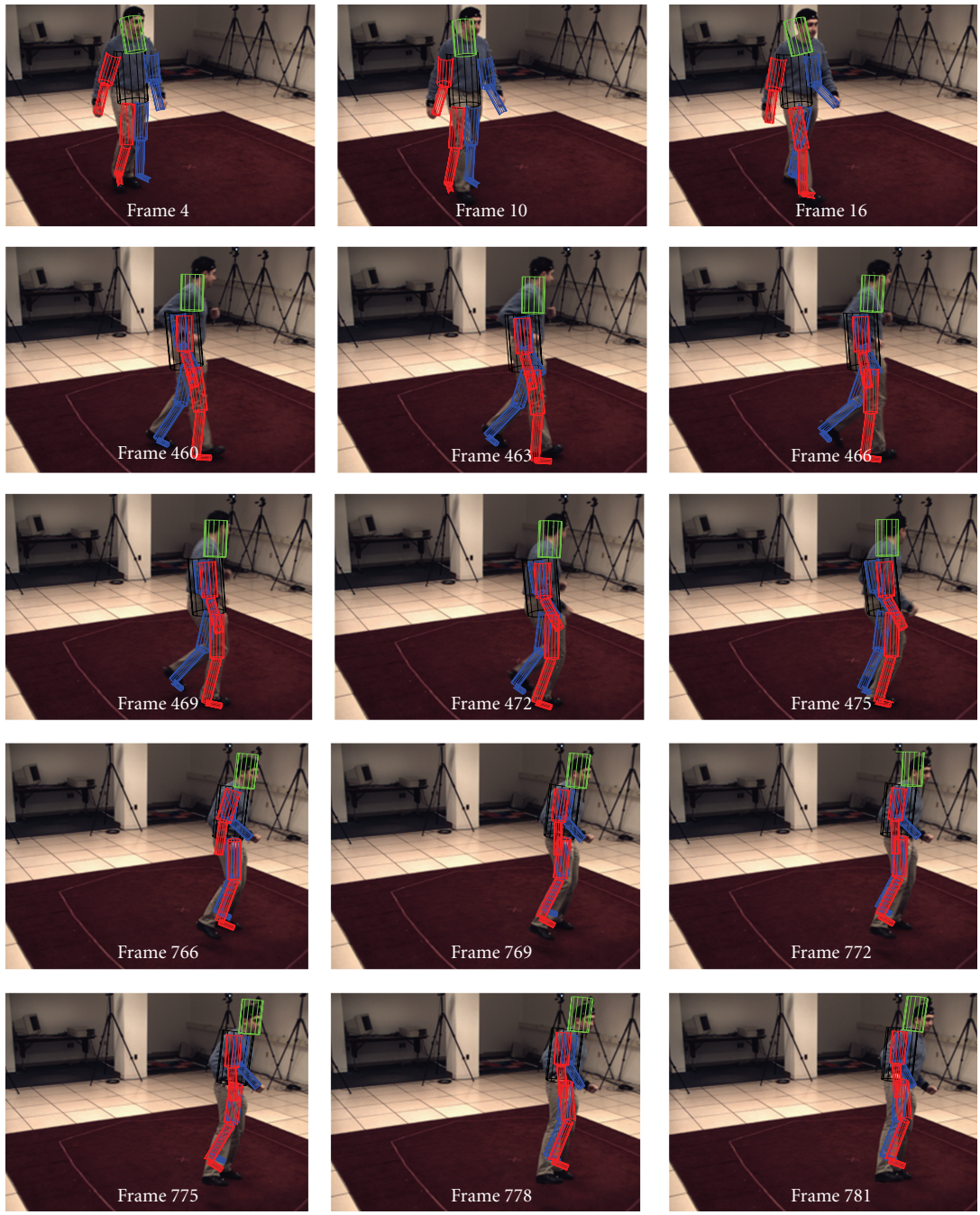


Figure 1: Tracking results for the HumanEva *S2 Combo 1* camera C4 sequence. Recovered poses of walking (frames 4–16), jogging (frames 460–475) and balance (frames 766–781) are superimposed with the input image.

this Semantic MCM (SMCM) provides the behavioural semantics of a movement sequence.

2.3. Training Movement Clusters. The learning process shown in Figure 4 proceeds from a set of of movements that are acquired from an MOCAP system as input data, to generate the dynamic MCM as output. Within the MCM, there is a statistical description for each MC, and for each transition between MCs.

Movements have high internal correlation, as they are made up from *consecutive*, therefore related, BFVs. Further, a single BFV has correlated parameters since body parts have synchronised movements. Consequently, we employ compression by principal component analysis (PCA) as the first stage to reduce the dimensionality of movements.

Clustering with expectation maximisation is employed to group similar $l_m + 1$ long compressed training movements. This results in a set of MCs. If a movement is separated into the first l_m and the last BFV then these define a

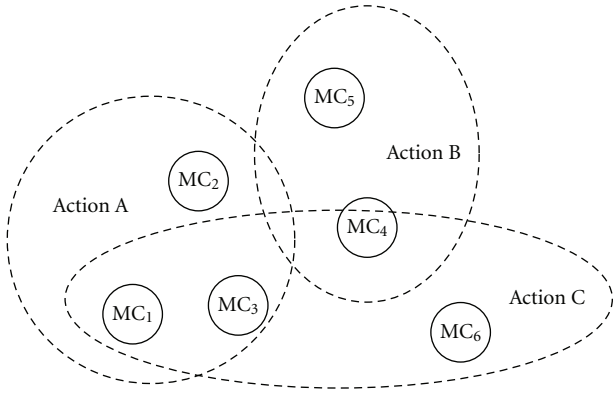


Figure 2: Movement clusters and actions. Action A results from any of the clusters $MC_1 \cdots MC_3$, Action B from MC_4 or MC_5, while Action C from MC_1, MC_3, MC_4 or MC_6. On the other hand a movement classified as MC_1 produces either Action A or Action C, etc., with a probability characteristic to MC_1.

Input: $\mathcal{M}$-the MCM
m-current movement
Output: b-generated BFV (NBFV)
(1) $b^* = {}_0m$ // the last BFV of the movement
(2) c =MC(m)
(3) $b \sim \mathcal{N}(b^*; \mathcal{M}.\mathcal{C}_c.NBFV.\mu, \mathcal{M}.\mathcal{C}_c.NBFV.\mathbf{P})$

Algorithm 1: NextBFV algorithm to generate the NBFV of the current movement.

movement to a new BFV transition. Therefore in the *cluster modelling phase*, for each MC, we compute a Gaussian distribution (i.e., the mean and covariance) of the PCA compressed l_m long movements, and a similar statistical representation of the *next BFV* (NBFV). The first defines the membership of an arbitrary movement within an MC, while the second provides the transitions to the NBFV, as suggested in Figure 5. This process is unsupervised and uses as input only the set of movements.

2.4. Movement Prediction. MCM allows pose prediction if the current movement is known. This prediction is a necessary step for generative tracking algorithms, such as the HPPF, but a synthetic motion sequence can also be used to verify if the model captures the correct dynamics of the human body.

Let us assume that the current movement, m, is known, with its last BFV, b^*. Algorithm 1 predicts the next BFV. First, the current movement's cluster is found (MC, line 2). Then, a new BFV is drawn from the learnt Gaussian distribution of the NBFV, with mean ($\mathcal{C}_c.NBFV.\mu$) and variance ($\mathcal{C}_c.NBFV.\mathbf{P}$) in line 3.

A repeated process of MC to MC transitions is illustrated in Figure 5. Since the current MC is known, the statistical model of each MC provides possible transitions to a next MC. This next MC is explicit. A new movement is created from circularly replacing the oldest BFV in the previous

Table 1: The set of MCMs from partition BFV of the whole PV. Each MCM has different complexity and refers to one or more body parts.

Model/level	Description
$\mathcal{M}_1$	Whole body
$\mathcal{M}_2$	Head
$\mathcal{M}_3$	Left full arm
$\mathcal{M}_4$	Right full arm
$\mathcal{M}_5$	Left full leg
$\mathcal{M}_6$	Right full leg
$\mathcal{M}_7$	Left upper arm
$\mathcal{M}_8$	Right upper arm
$\mathcal{M}_9$	Left upper leg
$\mathcal{M}_{10}$	Right upper leg
$\mathcal{M}_{11}$	Left lower arm
$\mathcal{M}_{12}$	Right lower arm
$\mathcal{M}_{13}$	Left lower leg
$\mathcal{M}_{14}$	Right lower leg

movement with the new BFV, b. Then, the new MC is that cluster which has the closest similarity to the new movement. Unlike an HMM, where the new state is explicitly defined by transition probabilities that can be drawn directly, in the MCM the transitions are hidden by the Gaussian model of the NBFV. However, compared to an HMM they estimate continuous parameters and define the transitions using movement, not just a single pose.

2.5. Model Parameters. The MCM has three parameters, which are the number of MCs, the length of a movement in video frames and the BFV components. A BFV can include either the full PV, or merely a subset. The latter is advantageous when considering the independence of limbs, each modelled with distinct MCMs. Table 1 defines 14 MCMs ($\mathcal{M}_i$) with different levels of detail: the full articulated body joint angle BFV (i.e., Whole body), the complete (Head, Left/Right Arm/Leg), the lower and upper limbs. Duplication (e.g., $\mathcal{M}_{10}$ is included in $\mathcal{M}_6$, while this is a part of $\mathcal{M}_1$) allows model specialisation for individual body parts.

We have observed that the head parameters are unstable, both in the training and tracking data. Therefore $\mathcal{M}_2$ is not used further in this study for prediction or analysis.

2.6. Movement Likelihood and Conditioned MC Probability. The probability of an arbitrary movement, m, from an MC is MC(m). To determine this value, we use the prior probability of the cluster $\mathcal{C}_i$,

$$\Pr(\mathcal{C}_i) = \mathcal{C}_i.\text{Prior}, \tag{2}$$

that is learnt during the training phase. To simplify the notation, the model $\mathcal{M}$ is implicit when referring to its clusters $\mathcal{C}_i$.

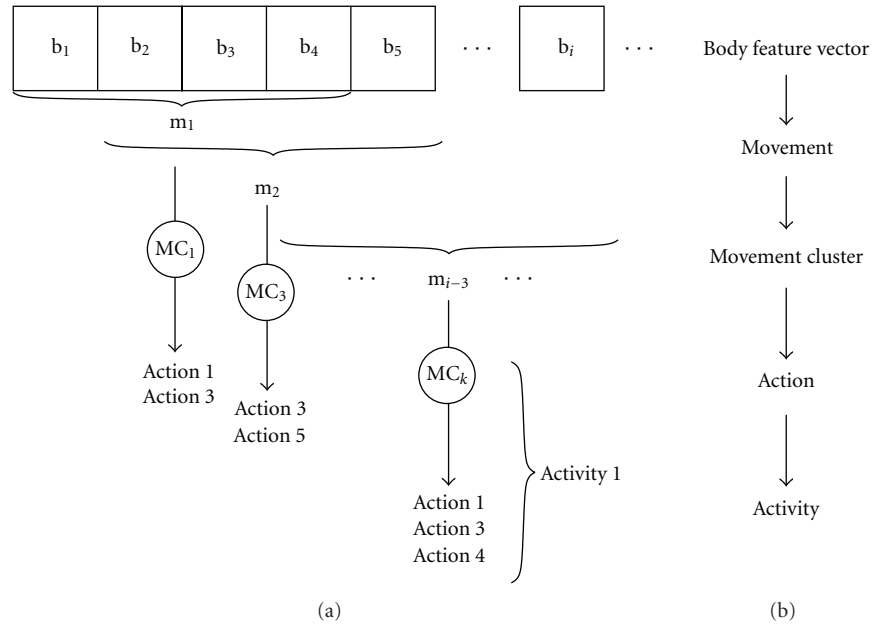


Figure 3: BFVs, movements, actions and activities. For an action primitive of length $l_m = 4$ body features $b_1, \ldots, b_4$ result in the movement m_1. The cluster MC_1, to which m_1 belongs, defines the possible actions (i.e., 1 and 3). Similarly, $b_2, \ldots, b_5$ define a different set of actions, *Actions 3* and *5*, by means of MC_3. Presence or absence of actions (over a time), or the coexistence of different actions in a temporally ordered manner, result in activities.

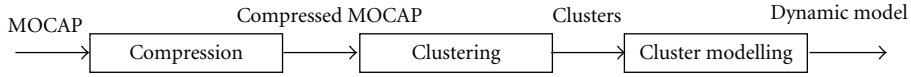


Figure 4: Motion model learning overview. First, the MOCAP training data is compressed to reduce the number of the correlated body parameters. Then, based on similarities, clusters are formed from the alike training data. Finally, features of these cluster are learnt.

With PCA compression the probability of a movement m conditioned by the MC is expressed by the probability of the compressed representation:

$$\Pr(\mathrm{m} \mid \mathcal{C}_i) = \Pr(\mathrm{pca}_\mathrm{m} \mid \mathcal{C}_i). \tag{3}$$

A normal probability density function models the cluster of the compressed movements, therefore

$$\Pr(\mathrm{m} \mid \mathcal{C}_i) = c_t \cdot e^{\delta_{\mathcal{C}_c}^T(\mathrm{pca}_\mathrm{m}) \cdot (\mathcal{C}_c.PCA.\mathbf{P})^{-1} \cdot \delta_{\mathcal{C}_c}(\mathrm{pca}_\mathrm{m})}, \tag{4}$$

where $\delta_{\mathcal{C}_c}(\mathrm{pca}_\mathrm{m}) = \mathrm{pca}_\mathrm{m} - \mathcal{C}_c.PCA.\mu$ and c_t is a constant.

Further, the probability of a movement m being cluster $\mathcal{C}_c$ using Bayes rule is

$$\begin{aligned} \mathtt{Sim}_{\mathcal{C}_c}(\mathrm{m}) &= \Pr(\mathcal{C}_c \mid \mathrm{m}) \\ &= \frac{\Pr(\mathrm{m} \mid \mathcal{C}_c)\Pr(\mathcal{C}_c)}{\Pr(\mathrm{m})} \\ &= c_t \frac{1}{\Pr(\mathrm{m})} \cdot \mathcal{C}_i.\mathrm{Prior} \cdot e^{\rho}, \\ \rho &= \delta_{\mathcal{C}_c}^T\left(\mathrm{pca}_\mathrm{ap}\right) \cdot (\mathcal{C}_c.PCA.\mathbf{P})^{-1} \cdot \delta_{\mathcal{C}_c}\left(\mathrm{pca}_\mathrm{ap}\right). \end{aligned} \tag{5}$$

Finally, with maximum a posteriori likelihood, the MC of an arbitrary movement m is the most similar cluster:

$$\mathtt{MC}(\mathrm{m}) = \arg\max_{\mathcal{C}_c} \mathtt{Sim}_{\mathcal{C}_c}(\mathrm{m}). \tag{6}$$

2.7. Action Learning. MC formation is unsupervised, based on clustering similar movements, but they do not necessarily have semantics. If some movements are labelled, these are used to assign semantic labels to MCs, as now described.

The ground truth probability $P(\lambda \mid \mathrm{m})$ is one if a movement m has the action label λ, and zero if not. Since movements are defined over a duration, it is important to specify the time of reference for λ. Here, it is given by the *last* frame, that is, λ is defined by the previous and current BFVs.

The probability of a label λ conditioned by a movement cluster, $\mathcal{C}_i$, is the frequency of the label weighted by the movement similarity within the cluster:

$$\Pr(\lambda \mid \mathcal{C}_i) = \frac{\sum_\mathrm{m} \mathtt{Sim}_{\mathcal{C}_i}(\mathrm{m}) \cdot P(\lambda \mid \mathrm{m})}{\sum_\mathrm{m} \mathtt{Sim}_{\mathcal{C}_i}(\mathrm{m})}. \tag{7}$$

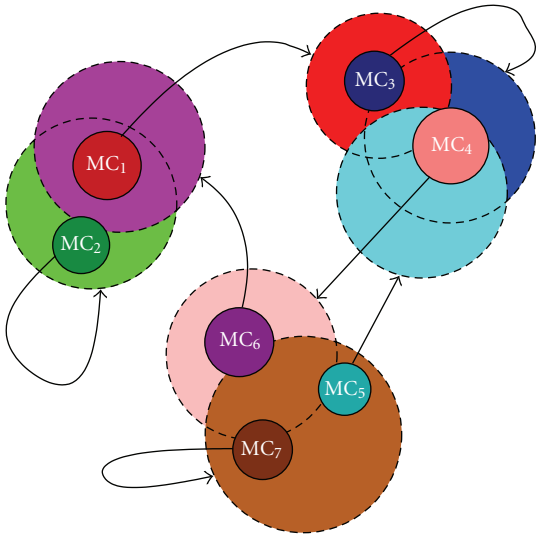


Figure 5: Visual example of an MCM. The MCs are represented by smaller continuous disks MC_i. The MC generates a new BFV suggested by larger, dotted circle and with all but the first BFV of the the current MC results in the new movement that is classified into an MC. MC_1 transforms into MC_3 or MC_4; MC_2 transforms into MC_1 or stays in the same state (but changes the parameter values); MC_3 to MC_3 or MC_4; MC_4 to MC_5, MC_6 or MC_7; MC_5 to MC_4; MC_6 to MC_1 or MC_2; and MC_7 to MC_5, MC_6 or MC_7.

The inferred probability of a label, given the current movement m, is the integrated marginal probabilities over all clusters:

$$\begin{aligned}\Pr(\lambda \mid \mathrm{m}) &= \sum_{\mathcal{C}_i} \Pr(\lambda, \mathcal{C}_i \mid \mathrm{m}) \\ &= \sum_{\mathcal{C}_i} \Pr(\lambda \mid \mathcal{C}_i, \mathrm{m}) \Pr(\mathcal{C}_i \mid \mathrm{m}) \\ &= \sum_{\mathcal{C}_i} \Pr(\lambda \mid \mathcal{C}_i) \Pr(\mathcal{C}_i \mid \mathrm{m}),\end{aligned} \tag{8}$$

with $\Pr(\lambda \mid \mathcal{C}_i, \mathrm{m}) = \Pr(\lambda \mid \mathcal{C}_i)$ from the functional dependence of the cluster $\mathcal{C}$ on the movement m.

A MCM with allocated movement similarities $\Pr(\lambda \mid \mathrm{m})$ is a Semantic Movement Cluster Model (SMCM). A SMCM can infer with (8) the probability of any action label, λ, given a movement, m.

2.8. Behavioural Analysis from Multihypothesis Tracking. When the input to behavioural analysis is from a multihypothesis tracker, such as the HPPF [2, 19], then the probability distribution of the AHHM recovered from the input video sequence is represented by the set of the particles, Ψ_t:

$$\Psi_t = \left\{ \mathrm{p}_t(i) \right\}_{i \in 1,\ldots,n_p}. \tag{9}$$

Tracking recovers the parameter distribution of the pose or the movement, which is used as input to the behavioural analysis. However, tracking benefits from the analysis of previous movement as inputs to the tracking process. There are two alternatives to how the PVs produce movements that are analysed.

First, since the particles of the HPPF can have a history in addition to the current PV, that is, they are movements, the movement distribution at time t becomes the current particle distribution:

$$\Psi_t^{\phi} = \left\{ \mathrm{p}_t^{\phi}(i) \right\}_{i \in 1,\ldots,n_p}. \tag{10}$$

Therefore, the label probability from (8) is computed as the expectation over the movement distribution, equal to the particle distribution. Hence,

$$\Pr\left(\lambda \mid \mathrm{m}_t^{\phi}\right) = \mathcal{E}_{i=1,\ldots,n_p} \left\langle \sum_{\mathcal{C}} \Pr(\lambda \mid \mathcal{C}) \Pr\left(\mathcal{C} \mid \mathrm{p}_t^{\phi}(i)\right) \right\rangle, \tag{11}$$

is the probability of the label λ.

With the second alternative, the movement is composed by conjoining the l_m consecutive current BFVs of the $\overline{p}_t$ estimated particle of the Ψ_τ. The current movement for the partition ϕ from the estimated particle is ${}_0\overline{\mathrm{p}}^{\phi}_{t-l_m+1}$, hence the current movement results in:

$$\mathrm{m}_t^{\phi} = \left[{}_0\overline{\mathrm{p}}^{\phi}_{t-l_m+1}, {}_0\overline{\mathrm{p}}^{\phi}_{t-l_m+2}, \ldots, {}_0\overline{\mathrm{p}}^{\phi}_{t} \right]. \tag{12}$$

This with (8) and (9) define completely the probability of label λ.

While formulation (11) requires the same MCM to be used for tracking and analysis, the latter allows use of an MCM with different l_m and $n_{\mathcal{C}}$, independent of the MCM used for tracking.

3. Analysis of a Trained MCM

In this section, to validate the MCM methodology, we analyse a trained MCM, and predict the transitions between MCs. For this, the MOCAP data of the HumanEva datasets [25] is used. The training videos and motion-capture (MOCAP) data provide ground truth on the limb positions in sequences of *Walk*, *Jog*, *Throw/Catch*, *Box* and *Gesture* activities. We follow the procedure described in Section 2.3.

3.1. Analysis of MCs. First, we analyse if the MCs resulted from the training are consistent semantically. For this, for each MC we show in Figures 6 and 7 the distributions of the constituent movements classed into the five categories *Walking, Jog, Box, Gesture, Throw and Catch*. For good classification, clusters should have similar movements arising from the same activity, while movements from different activities are expected to fall into different clusters. For example, in Figure 6(a) MC 13 is entirely composed of more than 2000 movements from the *Jog* activity. Conversely, MC 11 has in total 883 movements from the *Box*, *Throw/Catch* and *Jog* sequences. This cluster is the worst in Figure 6(a), since the membership of a movement in a cluster provides the least constraint on the activity of which the motion is a member. This artefact is motivated by the similarity of the *Box*, *Throw/Catch* and *Jog* sequences. For longer movements

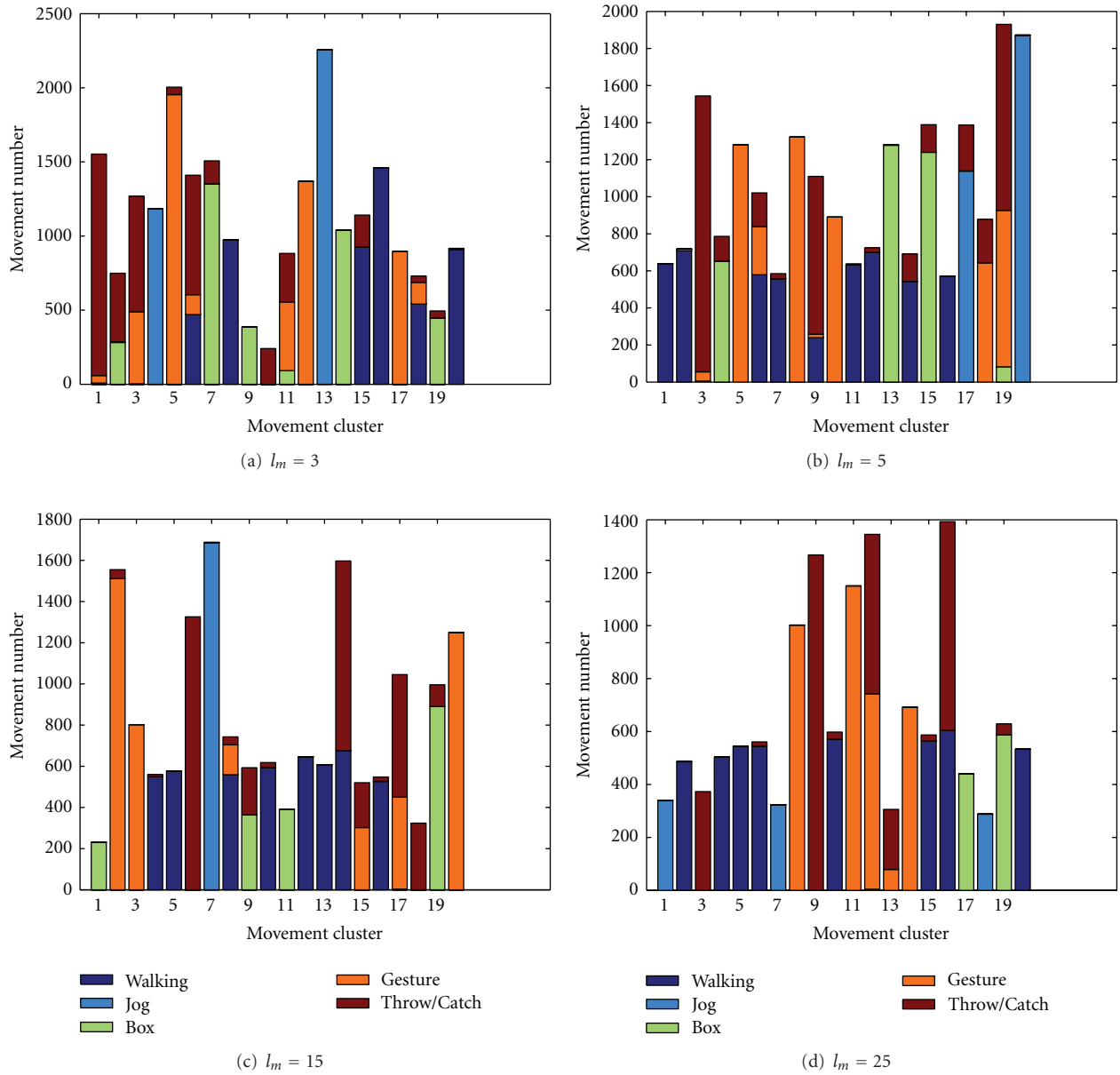


Figure 6: Movement length dependent cluster composition for the HumanEva basic activities, $n_{\mathcal{C}} = 20$ clusters and varying movement length l_m. The number of clusters with a mix of three activities reduces from 3 in (a) and (b), to 1 in (c) and 0 in (d).

having greater l_m, the clusters have more dominant activities. Thus, for $n_{\mathcal{C}} = 20$ clusters an increase in l_m results in better classification, since over a longer observation period the classification is more stable. Figure 7 shows the MC composition for $l_m = 25$ long movements for several cluster numbers. It suggests that with an increase of $n_{\mathcal{C}}$, the clusters with larger covariance are split and hence clusters are more discriminatory.

For objectivity, a measure of cluster uniformity is defined as

$$u = \sum_{\chi \in X} \frac{\left(c_\chi^{\max}\right)^2 / \sum_{\alpha \in A(\chi)} c_{\chi,\alpha}}{\sum_{\chi \in X} c_\chi^{\max}}, \tag{13}$$

with

$$c_\chi^{\max} = \max_{\alpha \in A(\chi)} \left(c_{\chi,\alpha}\right), \tag{14}$$

where X is the set of MCs, $A(\chi)$ is the set of activities of cluster χ, and $c_{\chi,\alpha}$ is the histogram value of movements of activity α in bin (i.e., cluster) χ. The uniformity u is one if all clusters have movements from a single cluster only; otherwise it favours clusters with a higher number of movements, penalising those with fewer. The minimum value is $1/|A|$, the inverse of the cardinality of the activity set, for example, the minimum uniformity for the above 5 activities is 20%. This minimum occurs when all MCs, equal in number, are in a single bin.

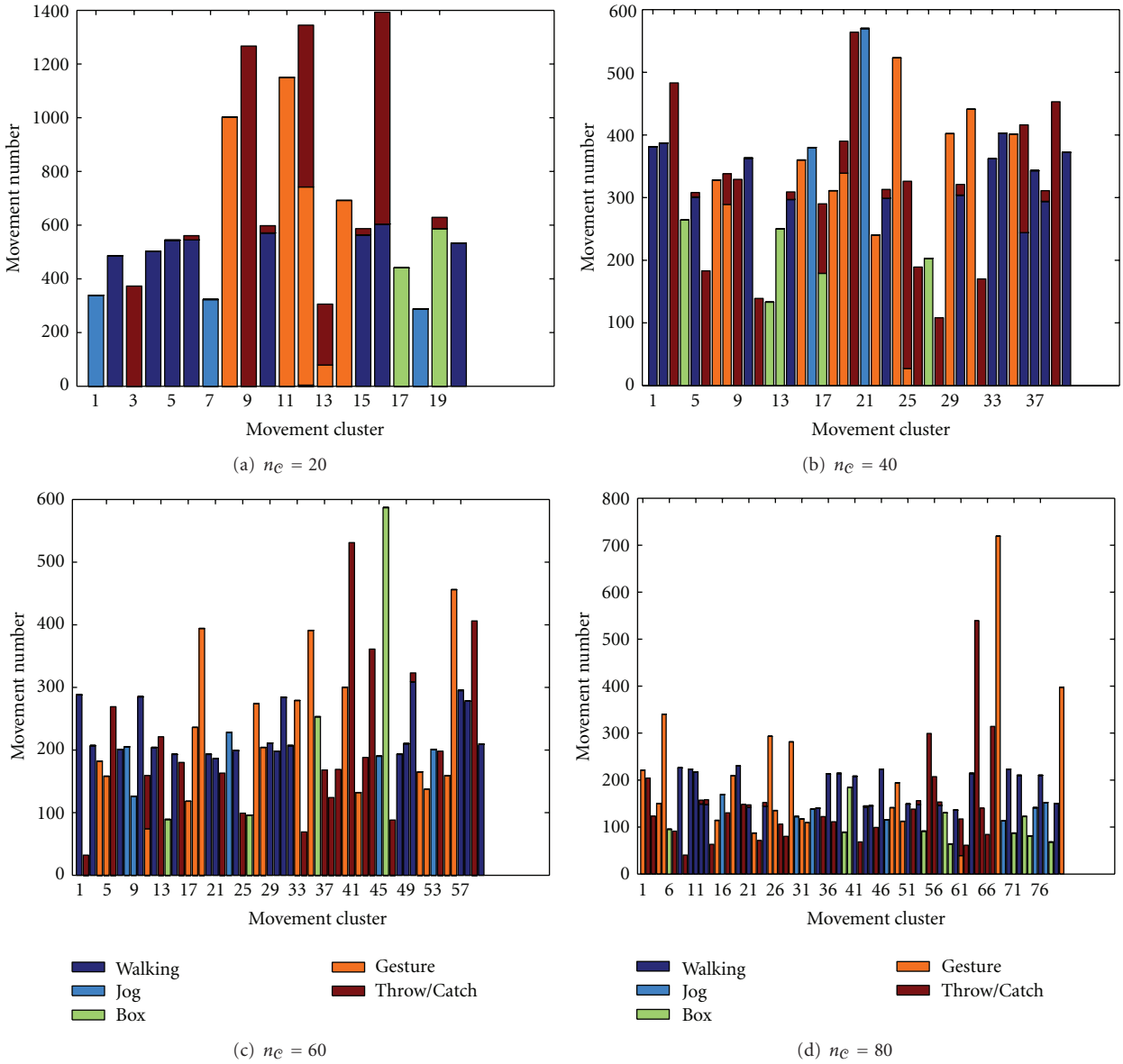


Figure 7: MC number-dependent cluster composition for the HumanEva basic activities, $l_m = 25$ movement length and varying cluster number n_c.

Table 2: Cluster uniformity u[%], in relation to the number of clusters n_c and the sequence length l_m.

n_c	l_m					
	1	3	5	15	25	35
20	91.5	91.7	90.4	91.5	93.0	95.7
40	94.6	93.6	94.9	94.5	97.3	97.4
60	96.2	95.7	96.7	96.9	99.5	99.1
80	96.4	96.8	97.6	98.5	99.3	99.0
100	96.5	97.4	97.9	98.8	99.6	99.4

Table 2 confirms that increases in both the movement length and the number of clusters enhance the MC uniformity. This is maximal for $n_c = 60$ and $l_m = 25$. Summarising, we can conclude that MCs contain similar movements and the discrimination between global actions is enhanced with more clusters and longer movements.

One would expect the number of clusters to be equal to the number of activities (i.e., five for HumanEva). This is not the case for many reasons: the high dimensional parameter space is multimodal, and clusters are used to classify not just one cluster of exclusive activities, but also actions, which can overlap and combine independently with other actions. This requires a value for n_c high enough to allow combinations between different actions. Adding more clusters is limited by the training set size; each cluster requires training to determine its mean and covariance with a number of member movements more than the dimensionality of the parameter space.

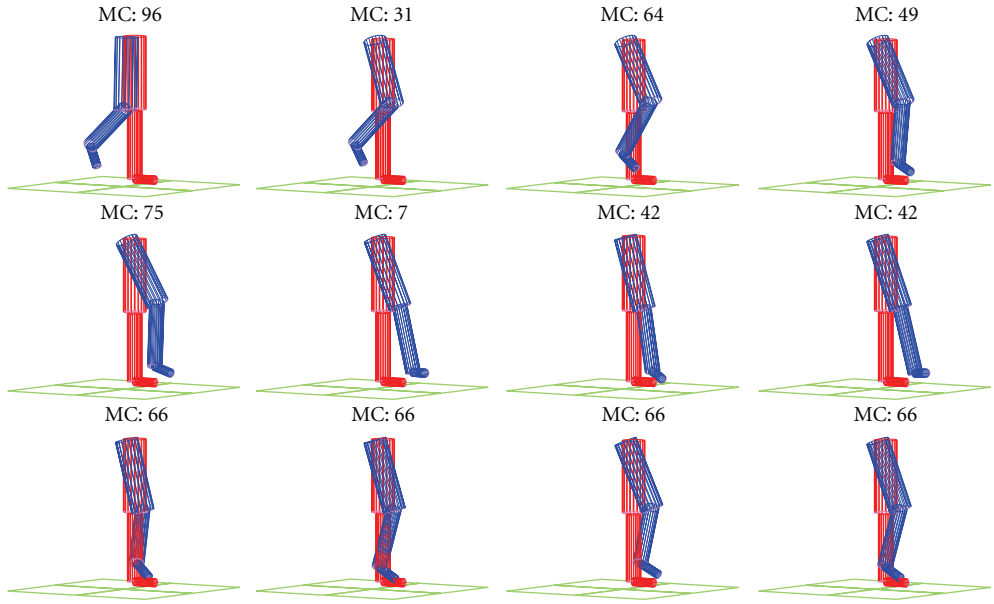


Figure 8: Left leg random motion (*mode* = *pose* only) with MCM $Model_5$. $n_C = 100$, $l_m = 3$. The initial MC is $mc_0 = 1$, not shown in the figure. This model is for the left leg only, therefore all other parameters are fixed.

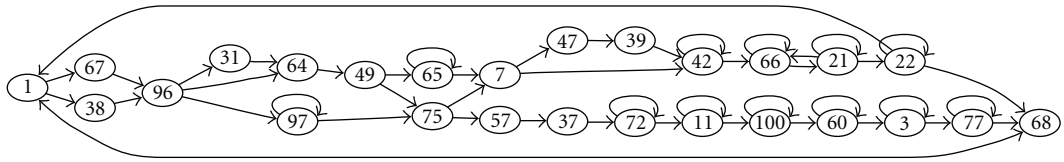


Figure 9: A leg MC transition sequence. The graph of 100 randomly generated transitions between leg MCs, with the first 12 poses shown in Figure 8.

3.2. Synthetic Motion Generation. An effective motion generator, Algorithm 1, with a good MCM should generate a natural motion. To demonstrate this, we show in Figure 8 a sequence of generated leg poses (i.e., BFVs) from a single initial pose using the left whole leg MCM (i.e., $\mathcal{M}_5$ from Table 1). It worth remarking that the corresponding MC transition diagram, Figure 9, is cyclic and allows different dynamics for the body part such as walking or other trained actions.

Transitions to the same MC are present, since consecutive poses are similar, but the stochastic component of the MC ensures that the poses are not identical. No interpenetration of body parts was observed for an MCM that includes multiple parts, however MCMs that model independent limbs only may allow this.

4. Semantic Analysis

Next, we evaluate our approach to behavioural analysis using the SMCM against ground truth and verification data from MOCAP poses of the HumanEva datasets.

4.1. Action in the HumanEva Dataset. The HumanEva dataset includes *Walk*, *Jog*, *Throw/Catch*, *Box* and *Gesture* activities. A movement is part of one of the five actions that directly corresponds to an activity that a video sequence represents. Therefore a movement can be explicitly described with this *global* label. Consequently, all 20 training sequences of HumanEva have been labelled for subjects S1 and S2 with one of the *Walk*, *Throw/Catch*, *Jog*, *Gesture* and *Box* global activity labels. Further to these global labels, five of the HumanEva *train* sequences were described by us with *detailed* labels from Table 3. As a result of this training phase, we can define a SMCM, $\mathcal{M}_1$, trained with labels from Table 3.

The global (e.g., *Walk*, *Throw/Catch*, etc.) and the detailed labels (e.g., *left stride back*, *left hand throw*) are considered at the same semantic *action* level. One can argue that global labels are activities, however components of a long sequence can be viewed as an action that defines the activity with the same name. Hence, *Walk* is an action and is part of the *Walk* activity. The *Walk* activity could also be inferred using detailed labels, but this is not the current aim. Further, a *Walk* action may be part of a more complex activity such as *Shopping*. Here, activity description is composed from concurrently detected global and local actions that provide detailed behavioural information.

4.2. Behavioural Analysis of HumanEva Dataset with the SMCM. Figures 10 and 11 show the behavioural probabilities of actions resulting from (8) for up to 190 movements per analysed HumanEva sequence.

Table 3: Detailed labels with descriptions and training sequences. For each left and right side, five pairs of action labels are trained with two or five HumanEva sequences.

Label	Description	Sequence
Left/right stride back	Left/right leg is moving forward from behind the right/left leg	S1 Walking 1, S1 Walking 3.
Left/right stride front	Left/right leg is moving forward ahead of the right/left leg	S1 Walking 1, S1 Walking 3.
Left/right arm forward	Left/right arm is moving forward	S1 Walking 1, S1 Walking 3, S1 ThrowCatch 1, S2 ThrowCatch 1, S2 ThrowCatch 3.
Left/right arm backward	Left/right arm is moving back-wards	S1 Walking 1, S1 Walking 3, S1 ThrowCatch 1, S2 ThrowCatch 1, S2 ThrowCatch 3.
Left/right hand throw	Right/left hand throw	S1 Walking 1, S1 Walking 3, S1 ThrowCatch 1, S2 ThrowCatch 1, S2 ThrowCatch 3.

The model parameters from Section 2.5, that is, sequence lengths l_m, cluster numbers $n_{\mathcal{C}}$, and BFV partition selection all affect recognition. The effect of these parameters are investigated next.

First, a whole body model, $\mathcal{M}_1$, was trained as described in the previous section, we then used the *validate* HumanEva sequences to assess the reliability of activity recognition on unseen video sequences. While subjects S1 and S2 have training data (from the *train* sequences, distinct from *validate*), subject S3 was not included in training. Hence, S3 evaluates recognition for an unseen subject. For each frame, that is, each movement ending at the current frame within S1 and S3, the label probabilities resulting from (8) are shown in Figure 10. The *Walk* label for the whole *S1 Walking 1* test sequence is well recognised; the four stride labels and the four arm forward and backward labels are observed with excellent periodicity, even though neither the training nor the analysis process was aware of this recurring nature. The least accurate recognition is for the *S3 Gesture 1* sequence, because of its similarities with the *Throw/Catch* activity.

The visual evaluation of sequence lengths l_m and cluster numbers $n_{\mathcal{C}}$ of the *S1 Walking 1* sequence from Figure 11 suggests that selection of sequence length is more important than the number of clusters. With higher sequence length, recognition of shorter actions degrades, especially for low cluster numbers. An increase of $n_{\mathcal{C}}$ results in finer detail. The transitions between detailed labels are smoother and have intermediate probability values.

Finally, Figure 12 shows the confusion matrices of the global actions, defined as

$$\mathrm{C}(\lambda_a, \lambda_b) = \mathcal{E}_{\mathrm{m}\in(\text{sequences of action } \lambda_b)} \langle P(\lambda_a \mid m) \rangle, \qquad (15)$$

the expectation of λ_a detected labels over all movements with ground truth label λ_b.

The figure represents the overall recognition performance in classifying all movements of the *validate* dataset sequences. Misclassification of *Throw/Catch*, *Gesture* and *Box* activities with *Gesture* or *Box* for $l_m = 15$, or no-detection for $(l_m = 25, n_{\mathcal{C}} = 100)$ are emphasised with longer sequence length.

Generally, the SMCM supplies independent action probabilities and therefore concurrently *Walk* and *Throw/Catch*, or even *Walk* and *Jog* can be recognised and their probabilities of actions do not sum up to one. In this test, we have exclusive actions, hence the lines of the confusion matrices are normalised to one. However, if none of the labels are recognised then all probabilities are zero.

4.3. Actions from Partial SMCMs. In the previous section, the full body BFV, that is, $\mathcal{M}_1$, was employed. However, all SMCMs (Table 1) provide action labels. Figure 13 shows the recognised label probabilities for the *S1 Walking 1* sequence for $n_{\mathcal{C}} = 100$ and $l_m = 3$ for the 13 SMCMs (all except the unreliable head MCM), and the *Overall* label with the overall average label probability of the SMCMs.

The response of the limb level models to a label that is defined by another limb confirms the strong dependence between the parameters. However, as expected, legs, and specifically whole legs (i.e., $\mathcal{M}_5$ and $\mathcal{M}_6$), best reproduce the periodicity of the motion. It is clear that models with fewer parameters are more specialised and provide finer probabilities of the detailed labels. Which model is most successful for action analysis has not been analysed extensively, although the averaged overall labels suggest that

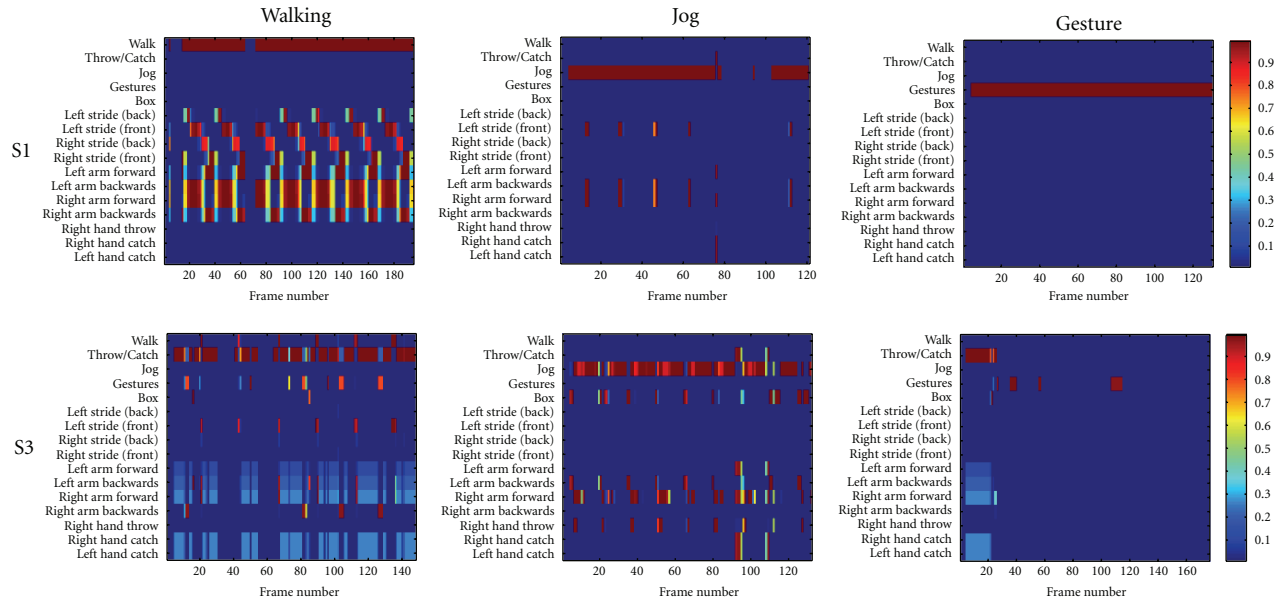


FIGURE 10: Recognition with known and unknown human subjects of *Walk*, *Jog*, and *Gesture* activities. Subject S1 was trained, while S3 was not. The model has $n_{\mathcal{C}} = 100$ and $l_m = 5$. The probability of labels (vertical) for each frame (on horizontal) is colour coded. For the first $l_m - 1$ frames no movement can be defined, and for frames 6–12 and 64–69 for *Walk* S1, frames 32–40 and 56–63 for *Walk* S3, frames 80–125 for *Jog* S1, frames 28–35, 41–55, 58–106 and 115–180 for *Gesture* S3 one of l_m the PV is missing, therefore no recognition is possible. This is visible by the vertical zero probability bands.

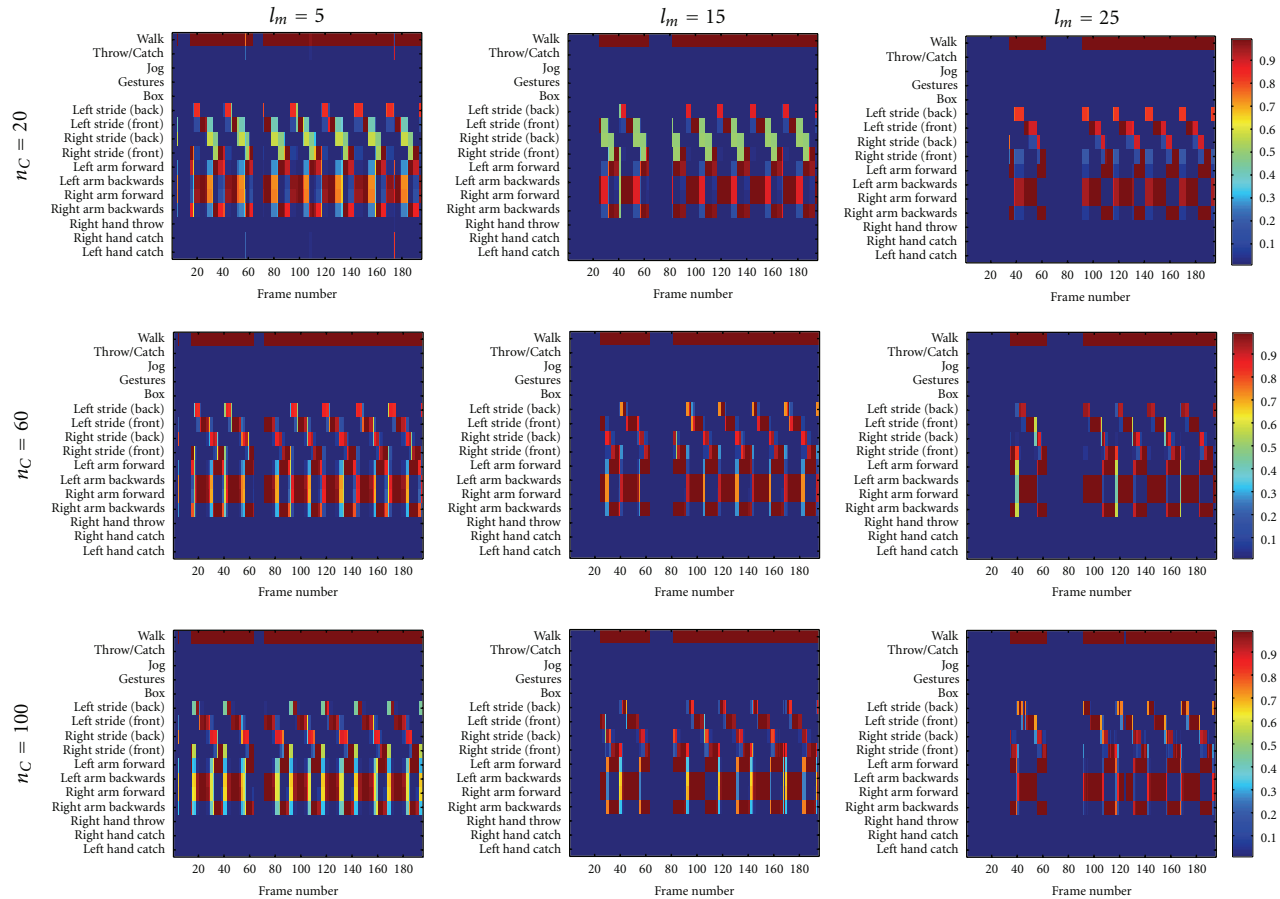


FIGURE 11: *S1 Walking 1* activity recognition for SMCMs with number of clusters $n_{\mathcal{C}} = 20, 60, 100$ and length of sequence $l_m = 5, 15, 25$. The zero probability vertical bands result from the missing pose information for frames 6–10 and 64–67. For these and the subsequent $l_m - 1$ frames a movement cannot be defined.

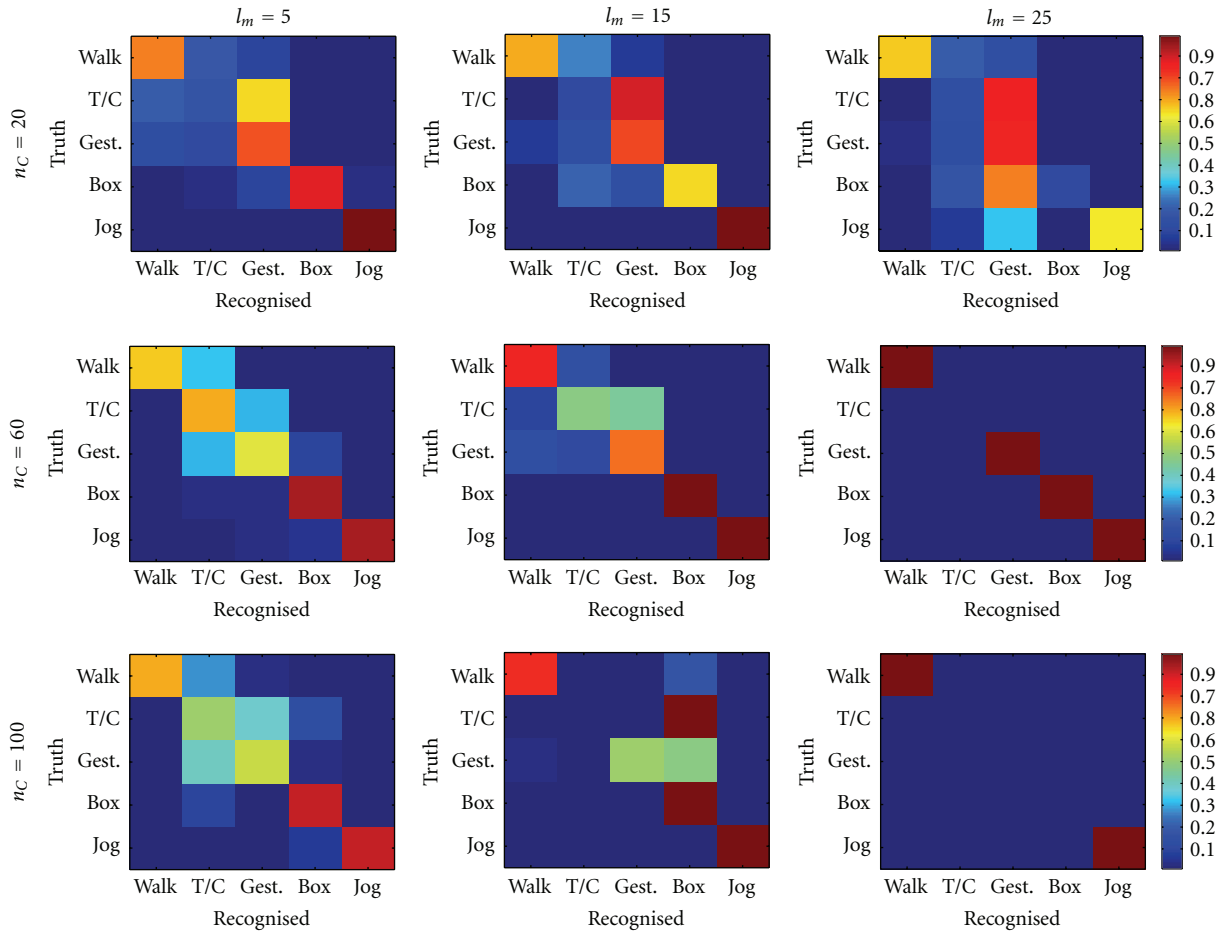


Figure 12: Confusion matrices for SMCMs with number of clusters $n_{\mathcal{C}} = 20, 60, 100$ and length of sequence $l_m = 5, 15, 25$.

combining SMCMs from a pool enhances label detection. However, the lower probabilities compared to individual SMCM probabilities convey that some do not contribute.

4.4. Recognition Sensitivity. To analyse the effects of parameter errors caused by an inaccurate BFV or movement recovery, the SMCM model was tested against white noise with $0 \leq \sigma \leq 2$ variance, added to ground truth (i.e., MOCAP) model parameters of the S1 and S2 subjects from the HumanEva-I dataset *validate* partition. The confusion matrices in Figure 14 show that recognition degrades with σ, and a shift towards *Box* and *Throw/Catch* actions that are similar to other activities, or have movements in common with them. This points out the necessity of accurate model recovery for good action recognition.

To evaluate the multiclass classification of actions, we have used either the full confusion matrix, or, for compactness, the ratio of the sum of the diagonal matrix to all matrix elements,

$$\zeta = \frac{\sum_i C_{i,i}}{\sum_{i,j} C_{i,j}}. \tag{16}$$

This gives a measure of classification recognition success rate, but is not a strict percentage because the sum of all action probabilities may be greater than one, that is, actions are not considered to be mutually exclusive.

The recognition success rate ζ is a function of the number of clusters and movement length, and depends on the added noise. For the full body SMCM, $\mathcal{M}_1$, the recognition success rates for several $n_{\mathcal{C}}$, l_m and σ are shown in Figure 15(a). An average tracking noise standard deviation of about 100mm corresponds to a $\sigma = 0.8$ [19, page 151], with the recognition success rates in Table 4. Figure 15(b) represents the recognition success rate variation on the right leg SMCM, $\mathcal{M}_6$. Both suggest a decrease in recognition success rate with increased noise. Similar to the results from Section 4.2, more MCs degrade recognition, especially for long movements. Figures show that error tolerance is best, 87%, for $l_m = 15$ and $n_{\mathcal{C}} = 60$ or $n_{\mathcal{C}} = 80$, recognition success rate being kept high for increased σ. While recognition success rate is higher for $l_m = 1$ the tolerance increases with l_m up to $l_m = 15$.

On the other hand, Figure 15(b) suggests that for a model with short BFV (i.e., right leg), the increase of either $n_{\mathcal{C}}$ or l_m improves recognition success rate subject to increased noise. Although ζ is lower, since independent limbs are less descriptive than the whole pose for global action detection, the drop with $\sigma = 2$ is less significant for $\mathcal{M}_6$, around 50% compared to 10%–0% of the error-free ζ with $\mathcal{M}_1$.

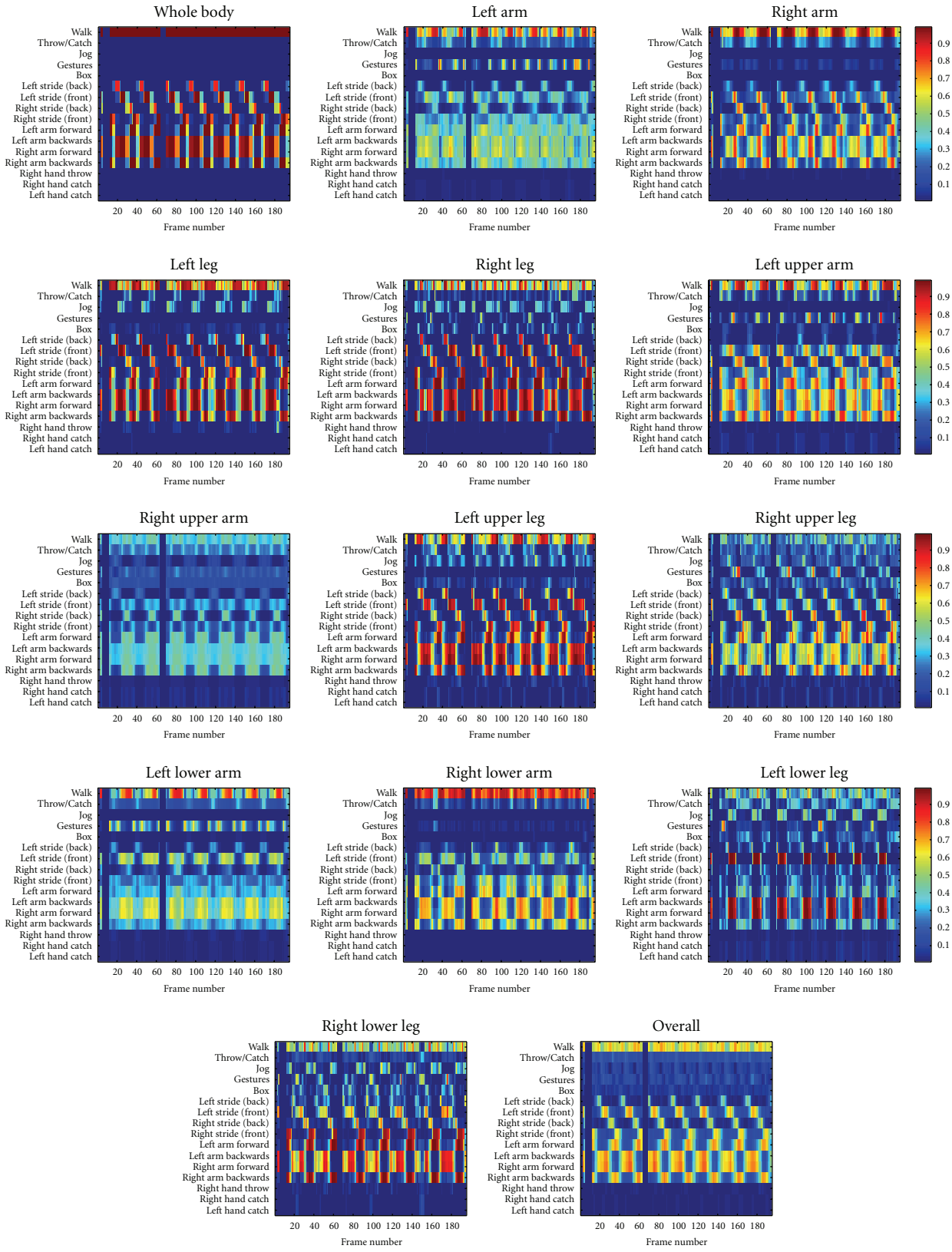


FIGURE 13: Recognition with SMCM set (Table 1, except head MCM) on the *S1 Walking 1* sequence. The last diagram shows the averaged overall performance of all SMCMs. The probability of labels (vertical) for each frame (on horizontal) is colour coded. Zero probability is shown for the first 2 frames, frames 6–12 and 64–69 with missing pose information in the $l_m = 3$ long movement.

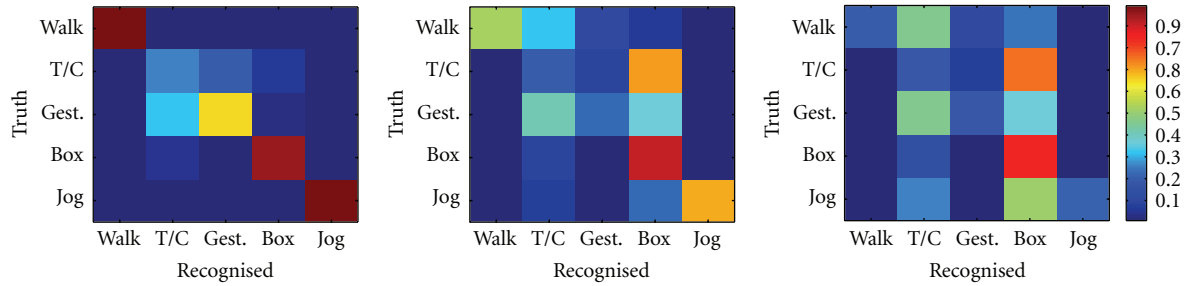
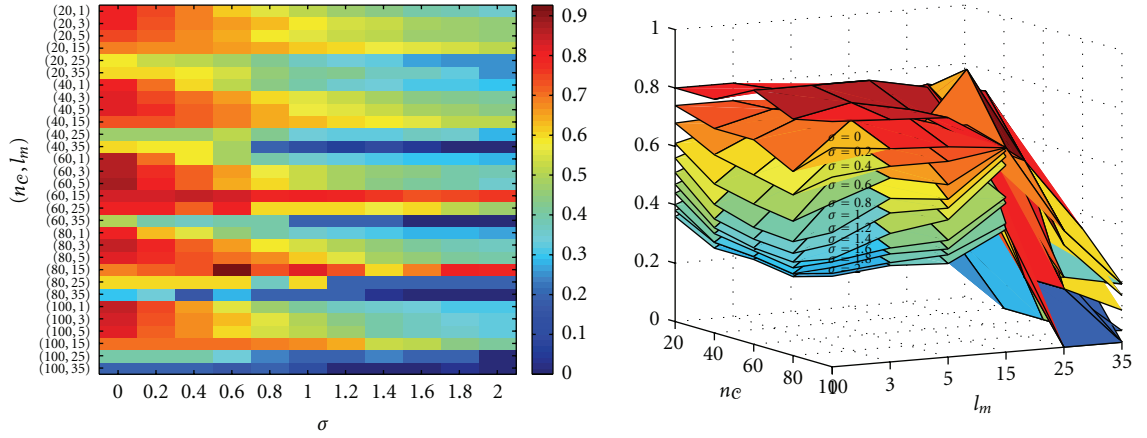
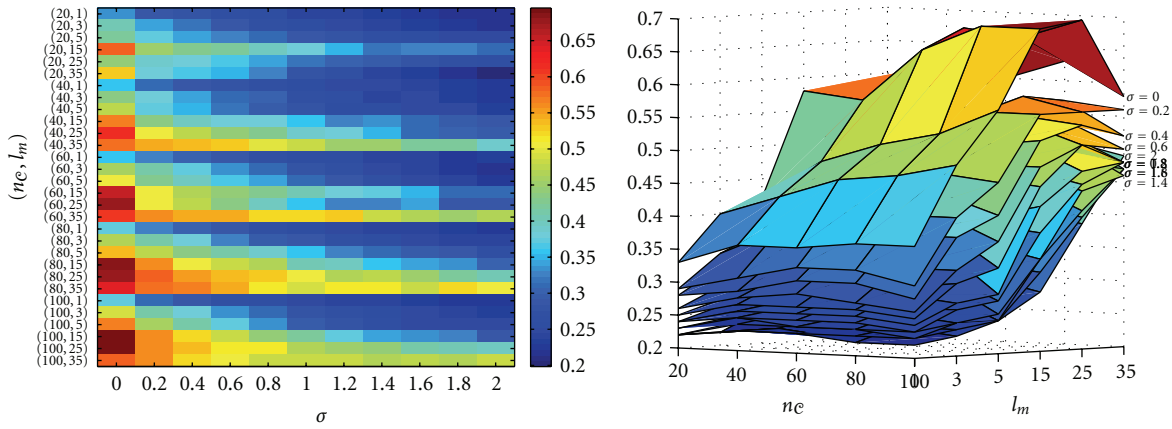


FIGURE 14: Action recognition with added noise. From left to right, confusion matrices for $\sigma \in \{0, 1, 2\}$ for a SMCM with $n_{\mathcal{C}} = 100$ and $l_m = 5$.



(a) Whole pose SMCM ($\mathcal{M}_1$)



(b) Right leg SMCM ($\mathcal{M}_6$)

FIGURE 15: Recognition success rate ζ variation for added noise $\sigma \in \{0.0, 0.2, \ldots, 2.0\}$. With increasing noise, the recognition is degrading.

4.5. Activities from Actions. Generally, activities are composed of multiple actions, however the SMCM classifies movement into actions only. The simplest inference of activity is the expectation of the activity labels over several, $N_{\text{activity}} \gg l_m$, frames and, in the extreme, over the whole sequence. The confusion matrices resulting from classifying each complete sequence (i.e., not each movement individually) are shown in Figure 16. The stationary activities are misclassified as *Gesture*, because of the similarities of the long, standing poses. Arguably, diverse short actions should be detected and assembled into activities with composition rules (fixed or learnt), for example, with a Stochastic Context

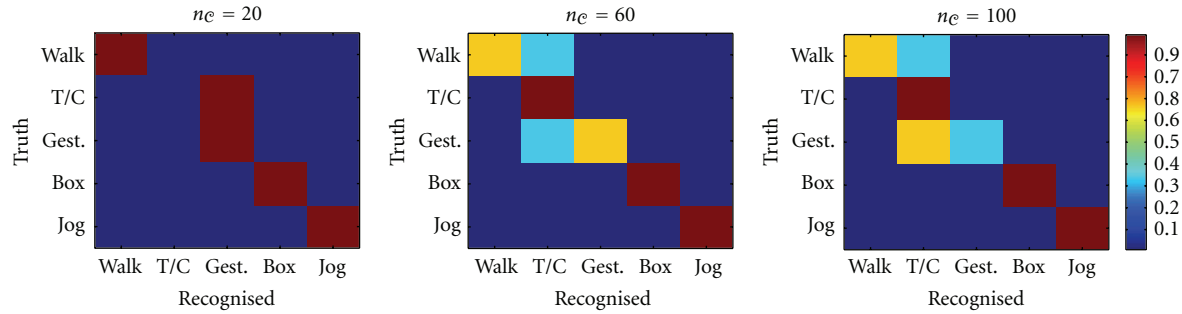


Figure 16: Confusion matrices of sequence classification. Each HumanEva sequence is classified into the activity which has the most corresponding action.

Table 4: Recognition success rate [%] ζ for $\sigma = 0.8$ noise of the analysed data in relation to the number of clusters $n_{\mathcal{C}}$ and the sequence length l_m.

l_m	$n_{\mathcal{C}}$				
	20	40	60	80	100
1	56	44	40	39	46
3	63	61	58	59	56
5	59	64	59	67	54
15	65	65	79	74	69
25	42	43	60	37	24
35	45	19	40	20	20

Table 5: Recognition success rate [%] with identical SMCMs for both tracking and behavioural analysis. $\mathcal{M}_1$ SMCM is used only for global action recognition.

l_m	$n_{\mathcal{C}}$				
	20	40	60	80	100
1	32	29	28	26	34
3	29	32	25	37	25
5	24	30	26	32	27
15	27	30	17	7	8
25	27	2	1	0	0
35	17	1	0	0	0

Free Grammar [26] as used by Ivanov and Bobick [27]. However, this raises questions about the robust recovery of the shorter, temporal labels.

5. Behavioural Analysis from Tracked Data

5.1. The Influence of the MCM Parameters. Finally, we evaluate behavioural analysis when the poses are provided by a tracking algorithm, the HPPF [2, 19], rather than the MOCAP data. The effects of $n_{\mathcal{C}}$ and l_m on the joint tracking-analysis system, connected with (11), are compared for 5×6, $n_{\mathcal{C}}$ and l_m values, Table 5, on the *Walk*, *Throw/Catch*, *Gesture* and *Jog* sequences. Resource limits restricted tracking tests to four out of the five HumanEva sequences, thus the confusion matrices are 4×5 with no *Box* activity. Recall from Section 4.2 that matrices do not normalise to one, since labels are considered independent, each with probability between zero and one. Null-lines are possible if no action cluster is recognised, that is, this is the nonrecognised action.

Random guessing one of four plus an *unknown* action (that includes *Box*) would result in a recognition success rate of 0.20 if these actions were mutually exclusive. However, as explained in Section 4.3 our recognition does not classify actions into one out of five categories, but computes a Bayesian probability that a sequence is an action, so that random guessing would in fact result in a value less than or equal to 0.2.

The success rates in Table 5 are greater for shorter movements. However, for greater l_m, none of the independent actions is recognised and the recognition success rate is zero. The recognition success rate decreases with l_m and $n_{\mathcal{C}}$. This was expected from Section 4.4 (Figure 15(a)).

5.2. Tracking and Analysis with Independent Models. With the formulation of (12), the SMCM parameters for behavioural analysis are independent of the SMCM used in the HPPF tracker for motion prediction. Therefore the dependence of semantic analysis by the SMCM on $n_{\mathcal{C}}$ and l_m is further analysed using the same HPPF-MCM tracker with the lowest errors ($l_m = 5$, $n_{\mathcal{C}} = 80$). The recognition success rates, shown in Table 6, are marginally poorer than for the HPPF integrated model, however they show the same degradation with both l_m and $n_{\mathcal{C}}$. As in Table 5, the recognition success rates are highest with $n_{\mathcal{C}} = 80$ and $l_m = 3$ or $l_m = 5$.

Comparing Tables 5 and 6, one concludes that only minor differences exist in the recognition success rates for identical respectively independent SMCMs used for tracking and for analysis. For both, the recognition success rates are low. Increasing either movement length l_m, or cluster number $n_{\mathcal{C}}$, degrades recognition. If both parameters are increased concurrently, there are improvements, limited however to $n_{\mathcal{C}} \leq 5$.

Table 6: Recognition success rate [%] for independent SMCMs for both tracking and behavioural analysis. $\mathcal{M}_1$ SMCM is used only for global action recognition.

l_m	n_c				
	20	40	60	80	100
1	30	34	31	34	35
3	31	23	28	35	26
5	27	29	22	31	33
15	29	32	25	8	4
25	23	3	0	0	0
35	18	1	0	0	0

5.3. The Influence of the SMCM Granularity. Analysis with the full pose SMCM, $\mathcal{M}_1$, leads to the question whether other models $\mathcal{M}_i$ from Table 1, with reduced BFVs, are better suitable. Therefore, recognition using the SMCM $\mathcal{M}_7$ (i.e., left upper arm parameters) has the recognition success rates shown in Table 7.

Comparing the recognition success rates of $\mathcal{M}_1$ and $\mathcal{M}_7$, Tables 5 and 7 suggest that a partition of parameters recognises actions better then the whole set; for models with smaller partitions (i.e., $\mathcal{M}_7$), a higher value of l_m results in better recognition. The first observation is motivated by the lower dimensional parameter space of the limb compared to the full pose SMCM (i.e., two against 18 dimensions). Since the longer MCs capture longer motion dynamics this explains the second observation. It is also observed that the effect of increased MC number is not visible using either of the models.

5.4. Recognition of HumanEva Sequences. Frame-by-frame analysis provides insight into our results, although it is subjective and qualitative. Figure 17 shows for the *S1 Walking 1* sequence the probability of the labels (horizontal axis) for each frame (vertical axis). The 13 diagrams for each sequence correspond to recognition with one of the $\mathcal{M}_i$ SMCMs. Both the tracking and the behavioural analysis use the same SMCM set, with $l_m = 5$ and $n_c = 80$.

The *S1 Walking 1* with the whole body SMCM recognises correctly the *Walk* action in the initial input, until frame 18, and in frames 45–72, while the other 90% of the frames have higher *Throw/Catch* probabilities. However, the diagrams show that six out of twelve local SMCMs (whole lower left and right arms, right upper arm, left upper left and right lower arm) produce high walking probabilities. The other SMCMs fail, and recognise *Throw/Catch* or *Box*.

In addition to the global action labels, local labels provide detailed description of the action. These are evaluated either visually, comparing them frame-by-frame to the image, or qualitatively, by their periodic alternation. The repetitive patterns of *Right stride back* and *front*, *Left stride back* (least visible) and *front* are best seen on the *Left upper leg* SMCM. The antiphase relationship of *left arm forward*, *right arm backwards* and *right arm forward*, *left arm backwards* is also visible in this diagram.

Table 7: Recognition success rate [%] with $\mathcal{M}_7$ SMCM is used only for global action recognition.

l_m	n_c				
	20	40	60	80	100
1	26	26	27	26	26
3	26	25	27	28	28
5	30	29	29	31	31
15	31	41	39	38	34
25	31	34	40	36	35
35	35	38	33	46	42

Figure 18 shows the labels recovered from the start of the *S1 Walking 1* sequence using the $\mathcal{M}_1$ SMCM superimposed on the input *S1 Walking 1* sequence frames. For clarity, labels are grouped into *General*, *Arm (left/right)* and *Leg (left/right)* semantics. Only labels with probability above 0.5 are displayed, in blue, and those with above 0.8, in green. The *Walk* action is recognised in 11 frames, while it is misclassified as *Throw/Catch* in 7 frames. Detailed labels are correctly detected without misclassifications, if detected, however in frames classified as *Throw/Catch*, they are missed. This was expected, since labels attached to MCs and *Throw/Catch* MCs, were not trained with arm and leg actions specific to *Walk*.

5.5. Tracking Quality. Comparing the results in this section to those of Section 4.4, which analysed the degradation in behavioural recognition success rate with noise, it is clear that the results are much poorer. Table 5 compared to Table 4 suggests a drop in the recognition success rate of about 20% for the tracked data compared to the MOCAP sequence.

This is due primarily to the uneven distribution of the errors for the different HumanEva sequences, which do not follow a well-behaved Gaussian distribution as with the simulations, but rather are prone to gross outlying errors caused by tracker failure on limbs in particular, a far from trivial task. For example, the lower-limb parameters are particularly error prone when compared to the upper-limb parameters. This suggests that more accurate and stable tracking would greatly improve the analysis with SMCM.

5.6. Recognition of the CAVIAR Sequence. The CAVIAR (EC Funded CAVIAR project/IST 2001 37540, found at http://homepages.inf.ed.ac.uk/rbf/CAVIAR) dataset, compared to the HumanEva videos, is representative of real CCTV that has increased interest for behavioural analysis. In these lower resolution videos with high perspective, the tracking is less stable and several frames have visual tracking errors. This jeopardises recognition of both whole body movements and longer movements. Similar to the HumanEva *Walk* sequence, the full pose SMCM is not effective in recovering the *Walk* actions. As for the *S1 Walking 1*, the *Left upper leg*, $\mathcal{M}_9$, SMCM provides the most detailed information about the visible periodic motion patterns, while seven levels recognise *Walk* with higher probability than other actions.

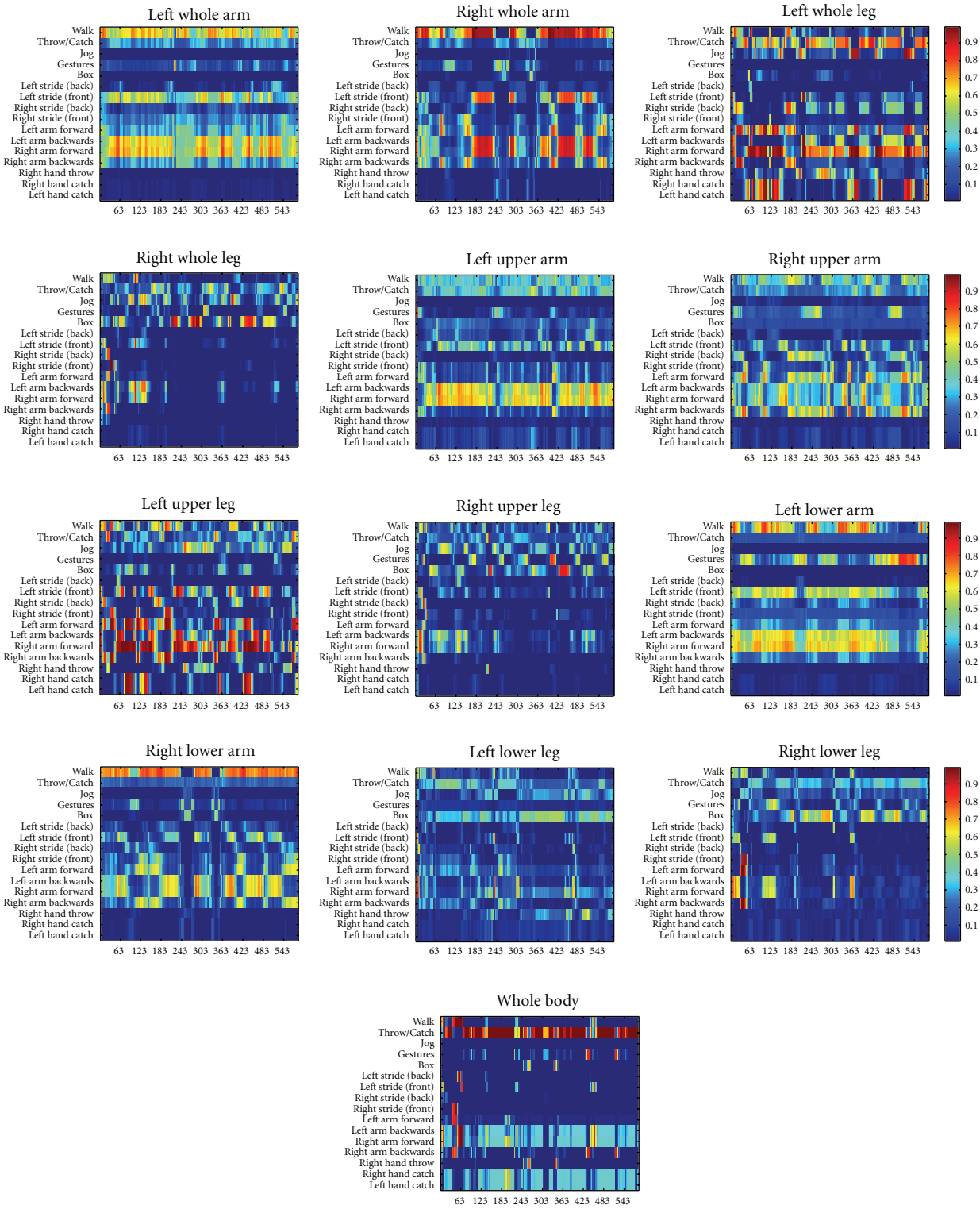


FIGURE 17: HumanEva *S1 Walking 1* sequence recognition. On each of the 13 MCM levels, for all frames (on horizontal) the probability of each label (vertical) is shown colour coded.

Since there are neither training data nor ground truth labeled activities, we use a SMCM trained with the HumanEva dataset and the results are evaluated visually. Figure 19 shows the action labels, superimposed on 6 out of the first 18 frames of the tracked sequence. Frames are misclassified by the similar but static *Throw/Catch* activity. However, the rest of the frames are classified as *Walk*, and include local action descriptions.

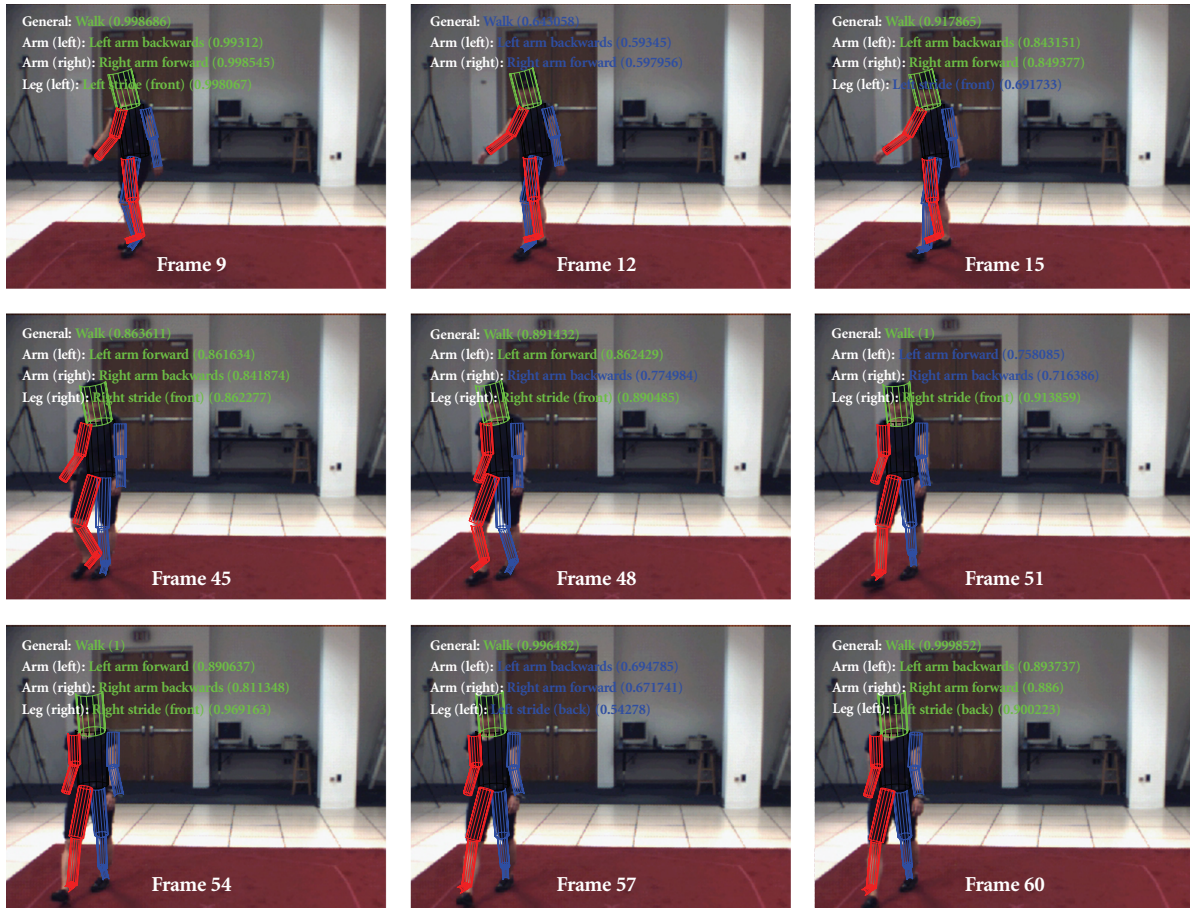


FIGURE 18: The recovered HumanEva *S1 Walking 1* labels superimposed with the input frames. Labels are manually grouped on different lines into *General, Arm (left/right)* and *Leg (left/right)* semantics. Only labels with probability above 0.5 are displayed, in blue, and those with above 0.8, in green.

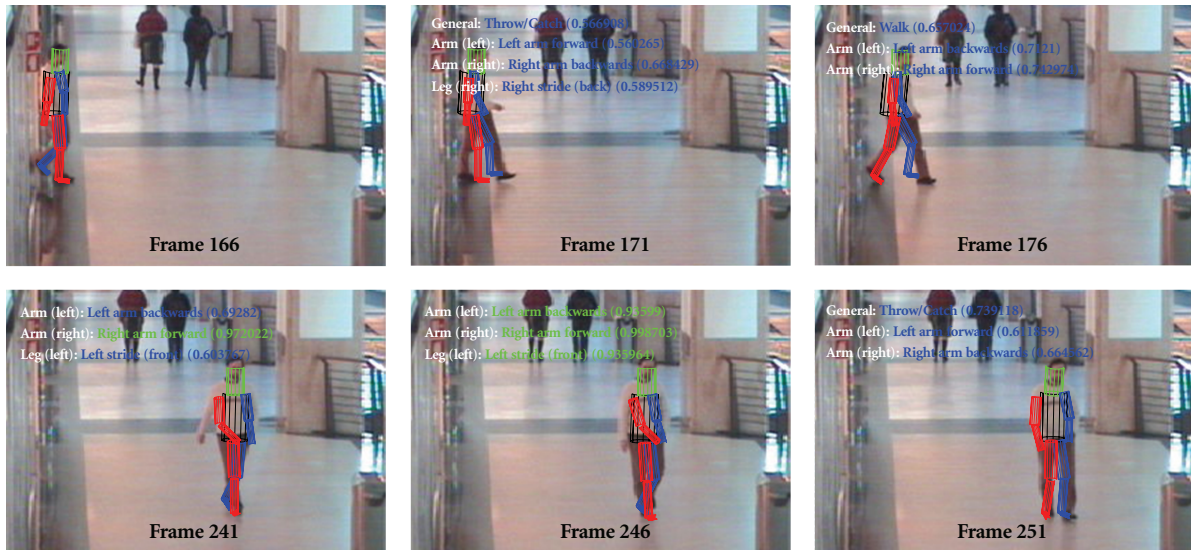


FIGURE 19: The recovered CAVIAR *EnterExitCrossingPaths1* labels superimposed with the input frames. Labels are manually grouped on different lines into *General, Arm (left/right)* and *Leg (left/right)* semantics. Only labels with probability above 0.5 are displayed, in blue, and those with above 0.8, in green.

6. Discussion and Comparison with Recent Work

Compared to trajectory analysis of the CAVIAR data [4] or on a football video [5], the SMCM infers complex behaviour patterns that need the dynamic analysis of body parts, rather than treating all movements as global trajectories on deformable regions. However SMCMs are restricted to a single individual, and we have not modelled multiperson interaction. In general, the SMCM would make such interaction very complex, for example to detect and describe the action of a boxer in parrying a blow using his left arm, made with the right arm of an opponent. Therefore, it may be more tractable in future to use individual limb movements and whole body trajectories to better model both individual and interactive behaviour respectively, as neither approach in isolation is effective in all situations. However, given the difficulty in extracting reliable limb dynamics in real data such as the CAVIAR sequence, it is understandable that much recent work has concentrated on whole body trajectories.

Another possibility is to find features that describe individual pose that might be more reliably extracted than the limbs, which have anatomical veracity but may be too difficult to extract for rapid progress. Action manifolds [8, 18] are similar to the MC, since they have a compact representation of actions. To compare directly their results with our own, we observe that the action detection normalised confidences of Gall et al. [18] are in the range of 20%–60% for the three action class of Walk, Jog and Balance sequences of the HumanEva; and in the range of 0% to 40% for the detailed, 10 action class of the TUM Kitchen dataset. The SMCM detection probabilities, as was shown in Figures 13 and 17, are independent and cover the full range valid of 0%–100% probabilities.

The SMCM is a multimodal action model, related to behavioral modeling with an HMM [4, 7] or to a switching model [17]. In this sense, the MCM is similar to the HGPLVM of Lawrence and Moore [10]. Darby et al. [11], Raskin et al. [12], Han et al. [13] do model the MOCAP data from the HumanEva datasets with HGPLVM as hierarchical latent subspaces, however they use this model for generative tracking and not behavioural analysis, so there is no possible direct comparison. In contrast to the MCM, the HGPLVM does not represent movement but is composed of static poses in 3D space.

Bo et al. [28] and Li et al. [29] use latent space only to predict motion. Compared to these, the SMCM does behavioural analysis as well as pose prediction, within a global parameter space, but also on the hierarchy of different levels of detail. This hierarchy provides the 13 MCMs for localised body part analysis.

Motion generation with MCMs is similar to Sidenbladh et al. [30] in the sense that both generate a new pose completing the previous pattern (i.e., a movement) with one new pose. However, the MCM is a compact, explicit model and in contrast to [30] does not search the whole set of training data. Our Gaussian state and transition model allows unseen data. On the other hand [30] is more accurate if memory usage is not critical and the pattern matches the training data well.

The SMCM is trained with a general set of movements for which semantic training can be independently and incrementally performed. This is preferred if semantic labelling is labour intensive.

Compared to the all above methods, SMCM may produce (as in Figure 18) more complex descriptions of human action, such as *walking [80%] with the right lower leg moving forward [70%]*. Each component action has an individual probability. Numerical comparison of recognition success rates for the *complete sequence* (Figure 16), with probabilities between 0% (e.g., *Throw/Catch* is recognised as *Gesture*) to 100%, is less well against many recent algorithms, such as Lui et al. [9] (60%–100% success), Yu and Aggarwal [7] (93.6% success), Han et al. [13] (55.3%–100% success). However the latter authors focus specifically on the classification of one action out of a limited set, with 3 to 8 actions. On the other hand, SMCM provides a frame by frame analysis, and it does operate concurrently as tracking data arrives. The overall classification is given by majority, which is unfair, for example when *Throw/Catch* and *Gesture* activities are distinct by only a subset of frames of the whole sequence.

We are not aware of any other algorithm that does both global and detailed activity analysis in parallel, independently, and with soft, probabilistic decisions. Arguably, this is highly desirable for composite and complex behavioural description.

7. Conclusions

We have developed a Movement Cluster Model (MCM) for modelling and prediction. We generated MCM models from an MOCAP sequence data by unsupervised clustering. We then used these training sequences to learn semantic labels corresponding to known activities, producing Semantic MCMs (SMCMs) for subsequent analysis of human behaviour in unseen sequences. Such evaluation shows that the SMCM achieves good recognition rates for general activity, as high as 87% given the best selection of movement length and number of clusters, and the periodic aspects of the repetitive actions are also well detected. This value is actually quite good, considering that activities are independent, possibly concurrent, from nonexclusive classes, and the classification is made instantaneously from a short movement vector not from the whole sequence.

MC uniformity tests suggest that longer movements and more clusters result in more individualised MCs, however increasing the length of a movement, or the number of movement clusters, has adverse effects. For low level, detailed actions, movement length should be short and modelled with many clusters, while for global actions longer movements with limited numbers of clusters are preferred.

The advantages of the SMCM are dual applicability for prediction and recognition; prediction of both periodic and

aperiodic actions; incremental addition or removal of action labels (as SMCM training is separated into an unsupervised and a supervised phase); inclusion of arbitrary features. The SMCM set defines a pool of probabilistic labels of both actions and simple activities with a medium duration. These could be used in the future for abstract symbolic analysis.

Complex activities require the temporal combination of multiple actions, each with extent. To simplify, activities are defined by one or a set of independent actions, and activity recognition becomes action recognition. The independent recovery of the trained action labels provides detailed information about the activity beyond simple classification into a set of limited actions.

Behavioural analysis by the SMCM allows modularity and flexibility, and abstraction from the input data, while maintaining a probabilistic modelling strategy. When movements are recovered by an articulated human body tracker, as opposed to the MOCAP system, SMCM based analysis provides detailed action symbols of the activity. The tests on HumanEva and the CAVIAR sequences show that a detailed description can be recovered. However, as evident from the confusion matrices, and anticipated by the tests on MOCAP data with noise added to the ground truth parameters, there are misclassifications of actions, or failure to classify them at all, due to model recovery errors in tracking. Hence, although the SMCM is effective, it does require reliable estimates of the pose from the video sequence to be used in normal CCTV analysis, and this is not trivial in relatively low-resolution video footage such as the CAVIAR data in particular. As expected, since good behavioural analysis requires good articulated tracking, performance was markedly better if multiple camera views were available.

Further, the approach is weak in classifying activities (i.e., whole sequences) by the majority vote of individual actions. This is because activities have a multitude of action components, and the most salient of these might not be the most frequent (e.g., *Throw/Catch* is best defined by the short throwing and catching action and not the most frequent standing). Moreover, similar movements, for example, all those that are classified as "standing", are part of different actions.

In this work, only articulated pose parameters were used for analysis. Positional or velocity parameters can be highly discriminative features of *Walk* and *Jog* and distinguish between static and moving activities. The BFV could include additional positional and velocity parameters. Behavioural analysis would almost certainly benefit from parameters that describe where and how fast the subject or subject limbs are moving, but the particles with larger dimensionality would increase tracking complexity.

Acknowledgment

Z. L. Husz would like to acknowledge the award of an Overseas Research Studentship that was essential for the performance of this work. The present address for Z. L. Husz is BAE Systems, Advanced Technology Centre, P.O. Box 5, Bristol BS34 7QW, UK.

References

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: a survey," *ACM Computing Surveys*, vol. 38, no. 4, pp. 334–352, 2006.

[2] Z. Husz, A. M. Wallace, and P. R. Green, "Evaluation of a hierarchical partitioned particle filter with action primitives," in *Proceedings of the 2nd Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHUM2 '07)*, IEEE, 2007.

[3] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[4] X. Ma, F. Bashir, A. A. Khokhar, and D. Schonfeld, "Event analysis based on multiple interactive motion trajectories," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 397–406, 2009.

[5] R. Li, R. Chellappa, and S. K. Zhou, "Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 2450–2457, 2009.

[6] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 185–205, 2005.

[7] E. Yu and J. K. Aggarwal, "Human action recognition with extremities as semantic posture representation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1–8, 2009.

[8] F. Nater, H. Grabner, and L. Van Gool, "Exploiting simple hierarchies for unsupervised human behavior analysis," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '10)*, pp. 2014–2021, 2010.

[9] Y. M. Lui, J. R. Beveridge, and M. Kirby, "Action classification on product manifolds," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '10)*, pp. 833–839, 2010.

[10] N. D. Lawrence and A. J. Moore, "Hierarchical Gaussian process latent variable models," in *Proceedings of the International Conference on Machine Learning (ICML '07)*, Z. Ghahramani, Ed., vol. 227 of *ACM International Conference Proceeding Series*, pp. 481–488, ACM, 2007.

[11] J. Darby, B. Li, and N. Costen, "Tracking human pose with multiple activity models," *Pattern Recognition*, vol. 43, no. 9, pp. 3042–3058, 2010.

[12] L. Raskin, M. Rudzsky, and E. Rivlin, "3D human body-part tracking and action classification using a hierarchical body model," in *Proceedings of the British Machine Vision Conference*, 2009, http://www.bmva.org/bmvc/2009/Papers/Paper085/Paper085.pdf.

[13] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia, "Discriminative human action recognition in the learned hierarchical manifold space," *Image and Vision Computing*, vol. 28, no. 5, pp. 836–849, 2010.

[14] A. F. Bobick, "Movement, activity and action: the role of knowledge in the perception of motion," *Philosophical Transactions of the Royal Society of London*, vol. 352, no. 1358, pp. 1257–1265, 1997.

[15] N. İkizler and D. Forsyth, "Searching video for complex activities with finite state models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '07)*, June 2007.

[16] R. D. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images—part I: a new framework for modeling human motion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 179–190, 2004.

[17] P. Turaga and R. Chellappa, "Locally time-invariant models of human activities using trajectories on the grassmannian," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 2435–2441, 2009.

[18] J. Gall, A. Yao, and L. van Gool, "2D action recognition serves 3D human pose," in *Proceedings of the European Conference on Computer Vision*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6316 of *Lecture Notes in Computer Science*, Springer, 2010.

[19] Z. L. Husz, *Articulated human tracking and behavioral analysis in video sequences*, Ph.D. thesis, Engineering and Physical Sciences, Heriot-Watt Unversity, 2008.

[20] W. Qu and D. Schonfeld, "Real-time decentralized articulated motion analysis and object tracking from videos," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2129–2138, 2007.

[21] L. Bo and C. Sminchisescu, "Structured output-associative regression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 2403–2410, 2009.

[22] I. Rius, J. Gonzàlez, J. Varona, and F. Xavier Roca, "Action-specific motion prior for efficient Bayesian 3D human body tracking," *Pattern Recognition*, vol. 42, no. 11, pp. 2907–2921, 2009.

[23] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, "Dynamical binary latent variable models for 3D human pose tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '10)*, pp. 631–638, 2010.

[24] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '97)*, pp. 568–574, 1997.

[25] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[26] A. Stolcke, "An efficient probabilistic context-free parsing algorithm that computes prefix probabilities," *Computational Linguistics*, vol. 21, no. 2, pp. 165–201, 1995.

[27] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852–872, 2000.

[28] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Fast algorithms for large scale conditional 3D prediction," in *Proceedings of the 26th IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1–8, 2008.

[29] R. Li, M.-H. Yang, S. Sclaroff, and T.-P. Tian, "Monocular tracking of 3D human motion with a coordinated mixture of factor analyzers," in *Proceedings of the European Conference on Computer Vision*, vol. 3952 of *Lecture Notes in Computer Science*, pp. 137–150, Springer, 2006.

[30] H. Sidenbladh, M. J. Black, and L. Sigal, "Implicit probabilistic models of human motion for synthesis and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV '02)*, vol. 2350 of *Lecture Notes in Computer Science*, pp. 784–800, Springer, 2002.